Date, Time & Location:	Integrative Cancer Research Face-to-Face Meeting #1 August 24 – 25, 2004 Bethesda Marriott Suites Hotel

Attendees:

Attendee	Affiliation
Alex Lash	Memorial Sloan Kettering
Ari Kahn	NCI
Arnie Miles	Georgetown University--Lombardi
Arumani Manisundaram	Booz-Allen-Hamilton
Baris Suzek	Georgetown University--Lombardi
Brian Gilman	Cold Spring Harbor/Panther Informatics
Carl Schaefer	NCICB
Cathy Wu	Georgetown University--Lombardi
Chris Kingsley	University of California, San Francisco
Christine Richardson	BAH/Kevric
Claire Zhu	Booz-Allen-Hamilton
Craig Street	University of Pennsylvania--Abramson
David Jewell	Dartmouth--Norris Cotton
David Kane	NCI/SRA
Don Baldwin	University of Pennsylvania--Abramson
Edith Zang	Institute for Cancer Prevention
Everett Zhou	University of North Carolina--Lineberger
Frank Hartel	NCICB
Harold Riethman	Wistar Institute
Hong Dang	Alpha-Gamma Technologies
Hongzhan Huang	Georgetown University--Lombardi
Jack London	Thomas Jefferson – Kimmel
Jennifer Brush	ScenPro
Jeremy Harbig	University of South Florida--H. Lee Moffitt
Jim Lyons-Weiler	University of Pittsburgh
John Powell	NCI
John Rux	Wistar Institute
Jomol Mathew	New York University
Juli Klemm	BAH/3rd Millennium
Kathleen Gundry	NCICB/SAIC
Kutbuddin Doctor	The Burnham Institute
Leslie Derr	NCICB
Lihua Zhu	Northwestern University--Robert H. Lurie
Liming Yang	NCI
Louise Showe	Wistar
Mark Adams	Booz-Allen-Hamilton

Mary Edgerton	Vanderbilt University--Ingram
Mervi Heiskanen	NCICB
Michael Keller	Booz-Allen-Hamilton
Michael Showe	Wistar
Patricia DiSandro	Booz-Allen-Hamilton
Patrick McConnell	Duke University
Paul Spellman	Lawrence Berkeley National Laboratory
Peter Covitz	NCICB
Phillip Stafford	Translational Genomics Research Institute
Preshant Shah	NCICB
Rakesh Nagarajan	Washington University–Siteman
Ram Chilukuri	NCICB/Oracle
Reechik Chatterjee	Booz-Allen-Hamilton
Rick Pense	Meyer L. Prentis--Karmanos
Sasikumar Thangaraj	NCICB/SAIC
Simon Lin	Duke University
Steve Marron	University of North Carolina--Lineberger
Steve Pouros	Panther Informatics
Stuart Fischer	Columbia University--Herbert Irving
Subha Madhaven	NCICB
Sue Dubman	NCICB
Tara Akhavan	NCICB/SAIC
Ted Liefeld	Massachusetts Institute of Technology Center for Cancer Research
Terrence Barrette	University of Michigan
Terry Braun	University of Iowa–Holden
Theo Wills	Booz-Allen-Hamilton
Tom Moloshok	Fox Chase
Tommie Curtis	NCICB/SAIC
Veena Rajaraman	Oregon Health and Science University
William Sanchez	NCICB/SAIC
Xiaoming Wang	University of Chicago
Yajun Yi	Vanderbilt University--Ingram
Yue Wang	Georgetown University - Lombardi

Agenda Items:

Day 1: Tuesday, August 24th 9:00 AM - 6:00 PM

I. Compatibility Guidelines

Speaker: Arumani Manisundaram (Booz Allen Hamilton)

Arumani gave an overview of the caBIG Compatibility Guidelines.

Items discussed during and after the presentation:

- The majority of projects within the ICR Workspace are targeting Silver-level compatibility within a year of starting their project.
- If the system is Silver in every way except n-tier, the system can still be

designated as Silver?

  - Yes, a system can theoretically be caBIG Silver-level compatible in a 1-tier system.

- The CTMS Workspace has a caBIG Compatibility Special Interest Group. They are looking to receive input from other participants in the caBIG community. It is possible this group may be moved to the Architecture Workspace.
- Compatibility and Best Practices need to be discussed separately,
- How will caBIG compliance be assessed and who will be doing this?
  - The cross-cutting workspaces are proposing to establish an independent body to determine level of compliance.
  - The CTMS caBIG Compatibility SIG has suggested that a validation suite could be developed.
  - It was stated that compatibility criteria be incorporated into the compatibility framework.
  - Amendments to the Compatibility Guidelines do no imply that software can become “decertified”.
- Is there going to be a software platform that is common within caBIG?
  - This is not necessary. What is important is that interfaces between applications are compatible. The caBIG Compatibility guidelines provide goals for compatibility but not details for how to achieve compatibility.
- Amendments to the caBIG Compatibility document will be made through the cross cutting workspace, with oversight from the Strategic Planning Working Group.
- What assistance are project teams going to have in order to achieve caBIG compatibility?
  - It would be useful to have cross-cutting participants working with each project team to provide guidance and assistance.
  - They would be able to red flag approaches that might ultimately interfere with Gold-level compatibility
  - Requirements and Specification Documents will ultimately get sign off by the cross-cutting workspaces but they should definitely be involved early on
- Is caBIG compatibility being considered in the review of NCI grants?
  - Not at this point. NCI at this point wants to make people aware of the effort.
  - Large scale programs (>1 lab) will want to look at the caBIG to see what can be incorporated.
  - When the 3 year pilot is completed and successful, you will likely see caBIG in more of the NCI grant solicitations. Until then, lots of flexibility.
- What has been the selection criteria for projects chosen for grid reference

implementations?

  - Projects have been identified that would be particularly useful for addressing specific questions regarding grid implementation that were raised during creation of the caGRID prototype.

II. Introduction to Model Driven Architecture

Speaker: Sashi Thangaraj

Sashi gave an overview of Model Driven Architecture and its benefits.

Items discussed during and after the presentation:

- What development tools does the NCICB recommend?
  - NCICB conducted an evaluation of available tools based on a list of defined criteria. Poseidon (gentleware.com) has been used up to this point.
  - Enterprise Architect was identified as a powerful UML modeling tool that is reasonably priced. This is the tool the NCICB will be using for caCORE development moving forward. NCICB is looking into making a subsidized version available to caBIG developers.
- Will the kit be platform specific to Windows?
  - It will be stepwise, some components will be Java.
- Does MDA define a testing or validation structure?
  - No. it does not. It is a framework. The use cases can be used to write scripts however, and can help develop unit test cases.
- At what point in the process is a data model used?
  - A data model can be used at the design phase.

III. Introduction to Unified Modeling Language

Speaker: Sashi Thangaraj

Sashi gave an overview of UML-based development and the artifacts relevant to this process.

Items discussed during and after the presentation:

- The NCICB's caCORE infrastructure can be seen as a resource for caBIG developers
  - This infrastructure can be used as a starting point. The intent is to offer a set of tools and approaches.
  - caBIG will not impose a development infrastructure for every group that wants to contribute to the grid.
  - A developer's toolkit is being developed and should be available by the end of the year. No name yet (caDev?).
- How do we incorporate legacy environments into caBIG?
  - One will need to retrofit data services and analytical tools to make that resource available and compatible with caBIG.
  - The caBIG Compatibility Document should be used as a

guideline for this new work.

- In terms of documenting current systems with UML, there are reverse engineering tools available to aid with this.

- What are the elements needed for "automatic code generation"?
  - 1) The model provides the structural description of the system
  - 2) A template is also provided that contains the programmatic logic/dynamic characteristics. The template dictates the programming language that is used for implementation.

IV. Object-to-Relational Mapping & Object-to XML Mapping

Speaker: Sashi Thangaraj

Sashi gave an overview of approaches and tools relevant to object mapping.

Items discussed during and after the presentation:

- If we have an existing system with a well defined relational model, why is it necessary to provide an object model?
  - It has been found, that in practice, that a relational schema and associated SQL has not provided the necessary vehicle for bioinformatics. It is a brittle structure that is easily broken when requirements are added to the system.
  - An object layer provides flexibility for change. Constant growth of a relational schema can cause bugs. An object layer may be best
- Is X linked capability in XML library part of or written for caBIO?
  - It is written specifically for caBIO.
- What is the relationship between more than one object layer?
  - It depends on the specific project.
- The XML binding tools are built specifically for caBIO. They are driven from a config file.

V. Practical Guide to CDEs and caDSR

Speaker: Tommie Curtis

Tommie gave an overview of common data elements: What they are, why they are useful, how they are managed

Items discussed during and after the presentation:

- In order for the Workspace to agree upon CDEs during development, a great deal of communication will have to take place. How will this be managed?
  - Agreed – this requires ongoing communication.
  - A listserv or forum may work well for this sort of activity.
- How much effort should be built into our Statements of Work for CDE development?
  - May need 1-2 people in each small group to be point people who play a coordinating role. For people playing this role, the level of

effort will be higher.

- In general, the level of effort will be project-dependent.

- One route to CDE creation is through UML modeling
  - A UML model represents a significant part of what will become metadata
  - A loader exists for loading a UML model into the caDSR
  - The UML loader is not currently a packaged tool. Currently, it requires the help of a developer to do the loading. Ultimately, the loader will be run by the Workspace Steward.
- What level of granularity should be considered when creating CDEs? How specific does one need to be?
  - One approach is to define a few key required attributes of a given data element and to make the rest optional.
- How do we keep track of who originally collected or manipulated the data?
  - If this is important to the research, this information should be tracked along with the other descriptors of the information.
- How will the process of CDE creation affect the workflow of developers? It seems that it would slow them down.
  - The ICR Workspace has flexibility in defining its process for CDE creation.
  - It is proposed that a workspace steward be assigned who is responsible for the bulk of this work.
  - Enhancements to the caDSR that shorten development time will be a priority
- What is the relationship between the caDSR and EVS?
  - The EVS system is linked to the caDSR and provides definition of the vocabularies/ontologies that should be used to contrain the values of the data elements.
  - CUI is a Concept Unique Identifier. This is how vocabularies/ontologies are linked to CDEs. This is currently a manual process.
  - Some types of structures can be manually build for semantic relationships among CDEs but the EVS system does a better job at that.
- Once you have your CDEs, what do you do with them?
  - They are used as constraints and definitions for an application that has data coming in and/or going out.
  - The data is then regularized as inputs and outputs and standardizes the data. Outbound data should all be represented as CDEs.
  - The idea is that when you set up a data management system, you keep track of what CDE goes with what data and you have

some means of displaying what CDE goes with what data.

  - Registering a model in the caDSR is the way of advertising and quering models in the form of CDEs.

- Can regular expressions be used to define permissible values?
  - Not currently, but this is an interesting idea.
- What is the difference between an object model and an ontology?
  - Object models are driven by use cases for a specific system.
  - An ontology modeling environment is used to provide information about data in general.
  - CDEs provide a linkage between object and ontologic modeling.
- Accessing EVS
  - The EVS provides an API which allows for synonym look-up.
  - The NCI thesaurus and metathesaurus are available through caBIO APIs.
  - The NCI thesaurus is published in OWL; the NCI metathesaurs is published in UMLS.

VI. Leveraging EVS in Applications
Speaker: Sashi Thangaraj

Sashi gave an overview of tools available for accessing the EVS in the caBIO architecture.

Items discussed during and after the presentation:

- When you see a term, how do you know which tree it is from?
  - It is up to the developer to configure the tree.
  - Need to request relationships built to support your application.
  - Note that the trees have "polyhierarchy" – a node can exist in more that one tree.
- How much data is in the NCI Thesarus?
  - There are 20 trees. The largest has about 7000 nodes, the smallest has a couple hundred nodes. The largest tree has 14 levels.
  - There are "is a" relationships from the lower nodes of the trees to the upper nodes.
- Inside NCI meta-thesaurus, all terminology is mapped to each other. One can search to see what terminologies are present and what concepts they represent.

VII. CDEs Harmonization
Speaker: Tommie Curtis

Tommie gave an overview of what harmonization is and how the process works.

Items discussed during and after the presentation:

- What are the differences between workflow and registration status?
  - Workflow shows how a CDE is progressing through an approval process.
  - Registration status defines a CDE for use across contents.

- Can well-established annotation be imported wholesale, to create a value domain such as the human gene names in HUGO?
  - Instead of importing whole list, use a reference standard
- How do you deal with standards that overlap and possibly contradict one another? Do you take subsets of standards and create a new standard?
  - There is not a straightforward answer to this. One would need to consider the specific application and make a determination on a data element basis.
- I need a new CDE. How long will it take to get it?
  - If it is very specific to the application and wouldn't be shared with other applications, it can take as little as an hour to create a CDE.
  - For a CDE that would be reused across applications, the process would take longer. At this point in caBIG, with projects just barely getting underway, we do not know how long harmonization will take.
  - One approach can be to propose a CDE during development and keep moving forward with development with the understanding that the CDE may need to be modified.
- Should there be an ICR project to caBIGify the human genome?
  - If the ICR Workspace decides this is important, this should be proposed.
  - Note that caBIO brings in the Santa Cruz human genome data through its DAS loader
- When does a new mapping of the human genome require CDEs to be updated?
  - If the new version has entirely new descriptive fields, then the CDEs will need to be modified.
- How frequently should you convert a UML model to a CDE? Since models can be regularly updated, how do you know when to reload.
  - It is important to have the concept of "release" in your application.

Day 2: Wednesday/August 25th/9:00 AM – 3:00 PM

Juli Klemm gave opening remarks on the day's agenda, which started with the external standards review, and was followed by SIG-specific discussions.

I. NCI External Standards Review

Speakers: Kathleen Gundry (SAIC) & Hong Dang (AGTI)

Kathleen Gundry gave an overview of the NCICB External Standards Document. Hong Dang gave a presentation of ICR-relevant standards from this study.

Items discussed during and after the presentation:

- Are there standards that allow querying of tissues/images for information such as gene expression?
  - There are standards for images, but don't know a lot about it. – refer to the External Standards Document.
- Hong Dang may be a resource to caBIG to help with curation/standardization of CDEs.

II. Genome Annotation SIG

Moderator/Speaker: Rakesh Nagarajan (Wash U)

Items discussed during and after the presentation:

- One of the biggest problems in genome annotation is how to identify the genes. What is the definition of genes? Can we agree on it?
- The group had a lively discussion of various gene identifiers and which ones are the best to use.
  - There are many gene identifiers. Each one is used for a different purpose. Therefore, which identifier to use when developing software is largely dependent on the context. There is clearly a need to associate identifiers with their context.
  - There maybe multiple CDEs for genes. LocusLink is an example of one CDE. In this case, LocusLink IDs are used as anchor points to join multiple sources of information.
  - From a developer standpoint, our software should interact with the databases in a way that ensures compatibility. It is desirable that the software is capable of interacting with a set of identifiers, instead of a single identifier.
  - There are mapping services available on the web that provide mapping between identifiers. Should caBIG incorporate such services (example: MatchMiner).
  - Some manual curation on the biologist part is unavoidable as ambiguity and inconsistencies exist among different identifiers.
  - It was suggested that the group look at what are the most commonly used identifiers, and decide on a set of identifiers that would cover 99% of the genes. LocusLink ID, UniGene ID, Genbank accession, Ensembl ID and NCBI ID are the ones the group proposed.
  - Not every gene fragment represents a unique gene, even though it may has a unique id. Locuslink is a possible solution.
  - One common identifier was suggested for all genes. But adding additional identifiers was considered a dangerous approach. It is best to use existing ones.
  - Should every identifier in the databases come from a CDE?

III. Proteomics

Moderator/Speaker: Simon Lin (Duke)

Simon gave an overview of SMOS, and standards relevant to proteomics.

Items discussed during and after the presentation:

- SMOS is being developed as a module within MIAPE that specifically deals with spectral data. It is focused on statistical modeling of the spectral data.
- Reproducibility of results is a major issue with statistical data analysis. Different statistical methods often produce completely different results even when applied to the same data. This calls for the need to store pre-processing and processing methods with the data.
- There is also the need for unbiased evaluation of data analysis results.

Submission of analysis methods together with the data should be made a requirement for journal publications.

IV. Microarray

Moderator/Speaker: Paul Spellman

Paul gave an overview of standards relevant to microarray repository. Paul is a member of the MGED consortium.

Items discussed during and after the presentation:

- Is MAGE-OM represented in the caDSR?
  - We (NCICB) are in the process of doing that. The approach to representing the MAGE ontology has not been finalized.
- Ultimately, microarray data has to be interconnected with proteomics LIMS, clinical data, RTPCR results, etc. How does MAGE deal with this?
  - Such efforts are important but are very much dependent on availability of funding, as this area overlaps significantly with other efforts such as PSI.
- Is there any effort on storing summarized results of microarray experiments, instead of raw data?
  - The Cancer Molecular Pages is one project that addresses that. Expression data are stored in association with genes and diseases so that one can query a gene for its expression across all cancer types.

V. Pathways Tools SIG

Moderator/Speaker: Carl Schaefer (NCICB), with Gary Bader (Sloan) standby on the phone.

Carl gave an overview of standards relevant to the pathways.

Items discussed during and after the presentation:

- PSI and BioPAX are perhaps most relevant to caBIG. SBML and CellML capture the dynamics but generally lack the capability of cross-referencing in a standard way.
- Are there standards being developed for naming pathways?
  - This is an important issue that is not currently being addressed.
- Is there a standard graphical way of representing pathways?
  - Not at this time.
- What about proprietary pathways such as those generated by Ingenuity? Will there be efforts to include these in the public domain?
  - There are many high-quality public sources available. Licensing of commercial pathways is not likely to happen due to high cost. However, as caBIG grows, its impact on the research community may become so great that commercial entities may choose to comply with caBIG standards.

- Is there a standard for pathway representation?
    - There is no overall standard on pathway representation.
    - One solution is to use graphic styles.

VI. Translational SIG

Moderator/Speaker: Terry Braun (U of Iowa, Holden)

Terry gave an overview of TrAPSS, which is a tool for mutation screening, for the purpose of discussing standards relevant to translational studies.

Items discussed during and after the presentation:

- What is the definition of translational?
    - The capability of bringing together information from genome annotation, pathways, etc., and applying it to clinical studies.
- Although caBIO, DAS exits, no standard and ontology fits every need. TrAPSS uses own standard and ontology.
- This SIG may want to investigate the Clinical Genomics Object Model.
- Important for this SIG to interact with the CTMS Workspace and the TBPT Workspace.

VII. Data Analysis & Statistical Tools
Moderator/Speaker: Veena Rajaraman

- Data analysis and statistical tools are consumers of the data represented in the other SIGs.
- Clinical Data Interchange Standard Consortium is pointed out as a starting place for clinical data integration.
- Need for standard for checking software algorithms, how data are handled.
- It was suggested that we should go beyond simply reporting on what algorithms and parameters are used, but the whole process of data analysis to ensure results are truly reproducible. GenePattern software addresses this need.

VIII. Summary:
Patrick McConnell (Duke), who is an Architecture liaison, called for submissions of use cases for:

1) Individual project needs.
2) Cross-project needs.

- It was suggested that instead of collecting use cases, the Architecture group should take all models available, and search for the solutions space.
- Focus on questions the researchers will have when using the system.
- Is there a web resource that lists all currently available use cases?

	o Look into the web forums, not just ICR, but other WS as well. • There is a data analysis resources list maintained by James Lyon-Weiler at UPMC. It would be very helpful if every component put up a use case along with the software tools. • Someone commented that the priority for caBIG should be getting the applications onto the Grid. Define things like, in order to get to the Grid, you must have X and Y, etc. • The question is then, who should go first? The next step should be assignment of pairs of ICR and Architecture developers.
Conclusions	• Participants are encouraged to communicate offline by themselves without moderation. SIG teleconferences could be used to summarize what comes off the offline communications. • Agenda for next SIG teleconferences • Assign people for decision making • Interlinking annotation sources across caBIG.

Action Items:

Name Responsible	Action Item	Date Due	Notes
Patrick McConnell	Create example use case for distribution within ICR	8/30/04	
All ICR Centers	Provide use cases to the Architecture WS	9/17/04	
Juli Klemm; Arumani Manisundaram; Christine Richardson; Mike Keller	Identify contacts within the cross-cutting workspaces to provide support to the ICR project teams	10/1/04	
Juli Klemm	Post meeting slides on the caBIG forum	8/30/04	
Juli Klemm; Claire Zhu; Reechik Chatterjee	Create and distribute meeting minutes	9/10/04	